



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Last Annotation of *Fugu rubripes* at JGI

A. Salamov, N. Putnam, A. Terry, I. Grigoriev, D.
Rokhsar, E. Loh, B. Venkatesh

April 19, 2006

Hinxton Fish Workshop
Cambridge, United Kingdom
November 24, 2005 through November 25, 2005

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Last annotation of *Fugu rubripes* at JGI

Asaf Salamov
DOE Joint Genome Institute

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

JGI Eukaryotic Genomics - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail News RSS Feeds

Address http://genome.jgi-psf.org/euk_home.html Go Links

EUKARYOTIC GENOMICS

Microbial Genomics Tree of Life

Home Current Releases Coming Soon Download Help

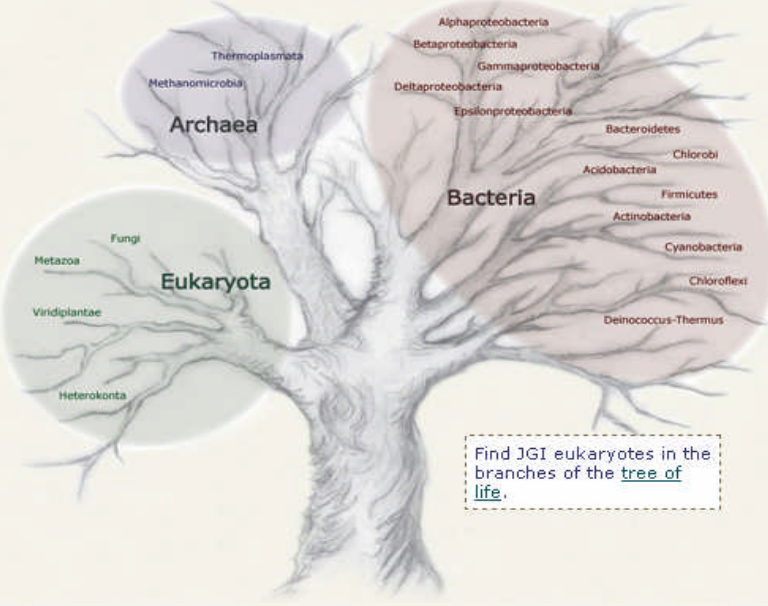
Select a JGI eukaryote and site location

From this site you can get details about our current and upcoming projects, or go directly to the [individual genome sites](#).

All of the individual sites include direct access to download sequence files, BLAST, search, view and navigate the genomic annotations.

For more information about how to use the site features and what tools are available, see the [online help](#).

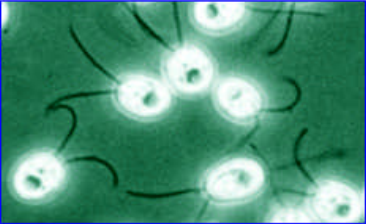
- Browse through our currently released [eukaryotic genome sites](#).
- Read the [online help](#) to find information about how to use the data, visualization and search tools.



Find JGI eukaryotes in the branches of the [tree of life](#).

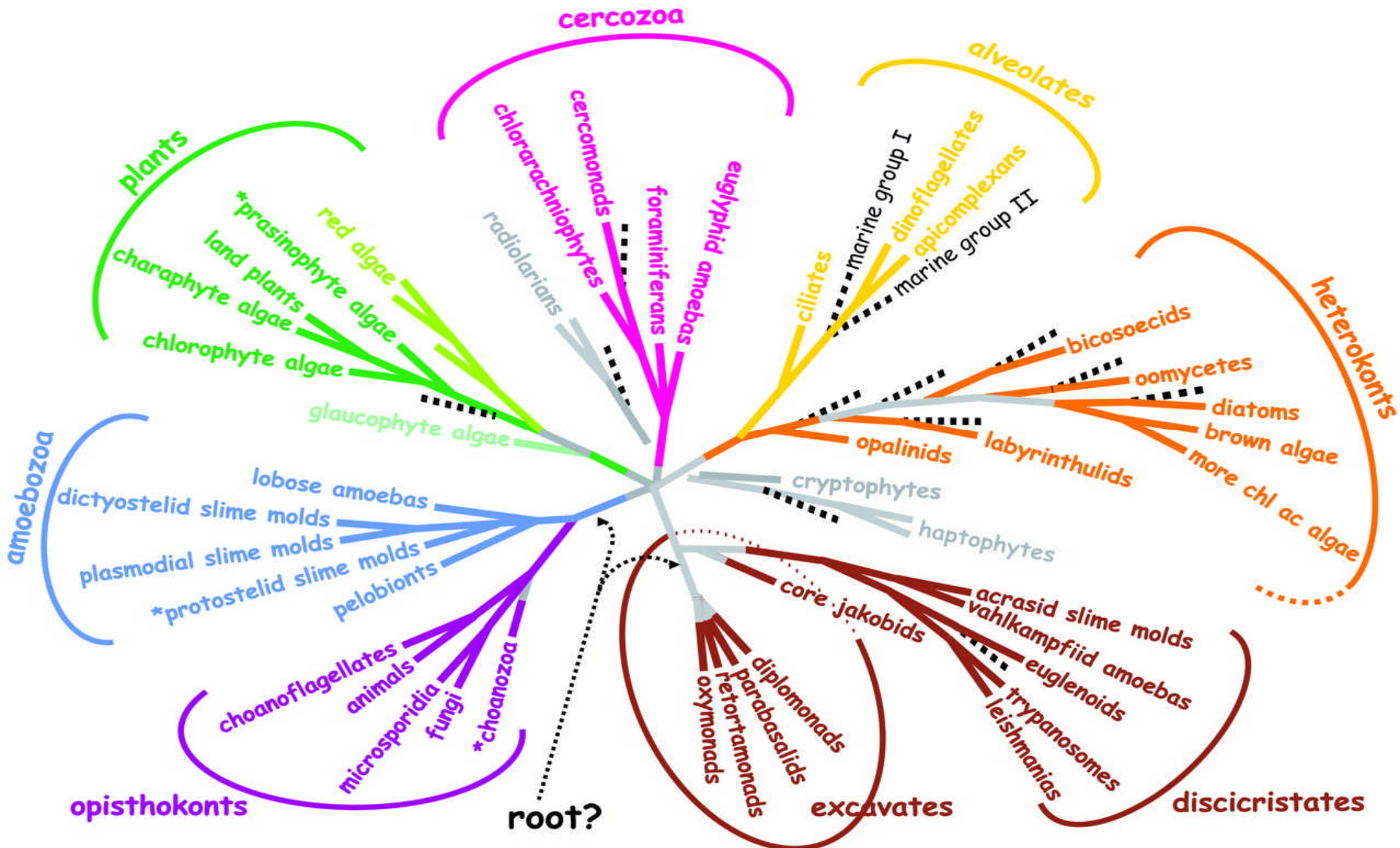
Eukaryotic Highlights

Chlamydomonas reinhardtii is a single celled chlorophyte. Highly adaptable, these green algae live in many different environments throughout the world. The relative adaptability and quick generation time has made *Chlamydomonas* an important model for biological research.

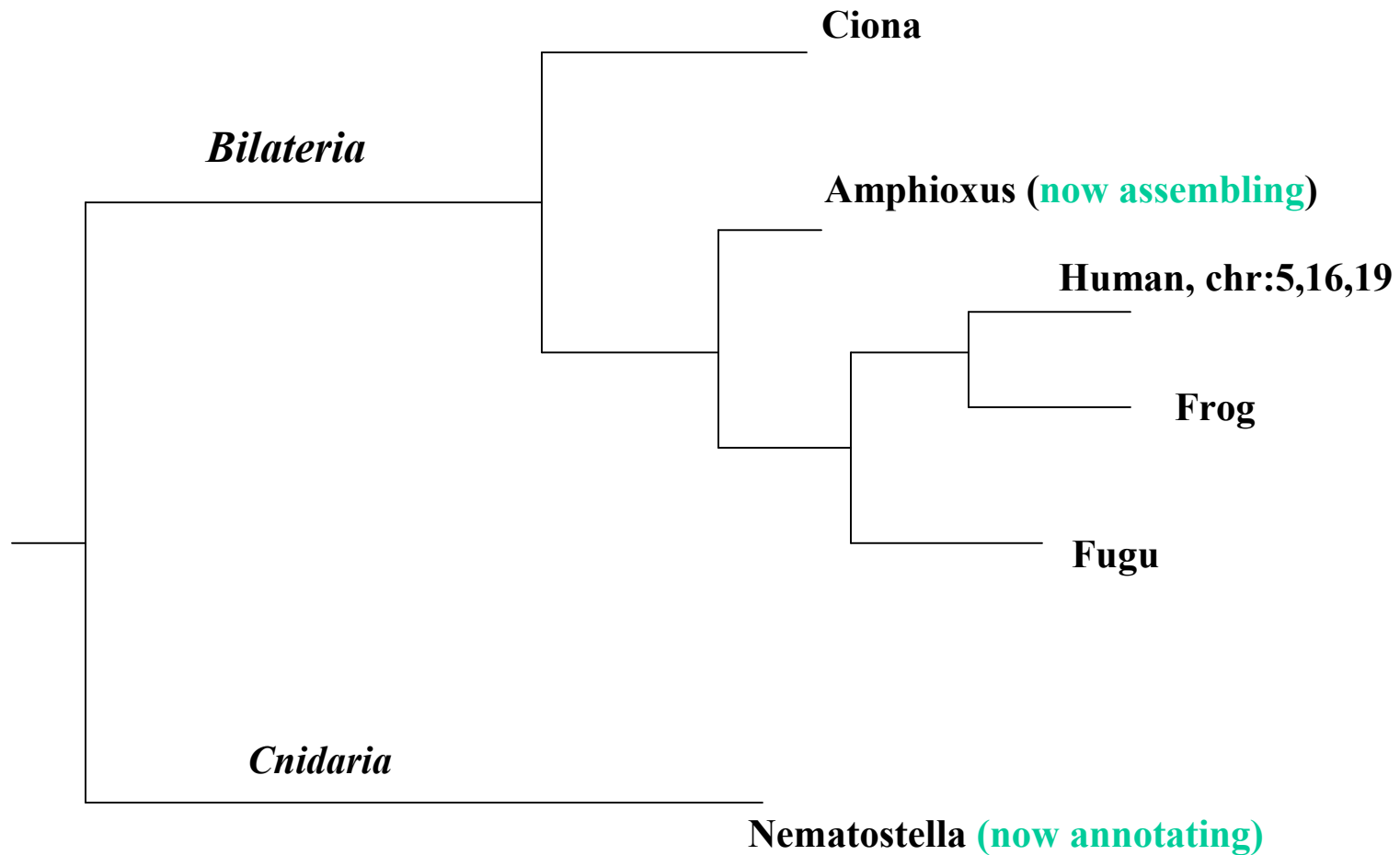


Internet

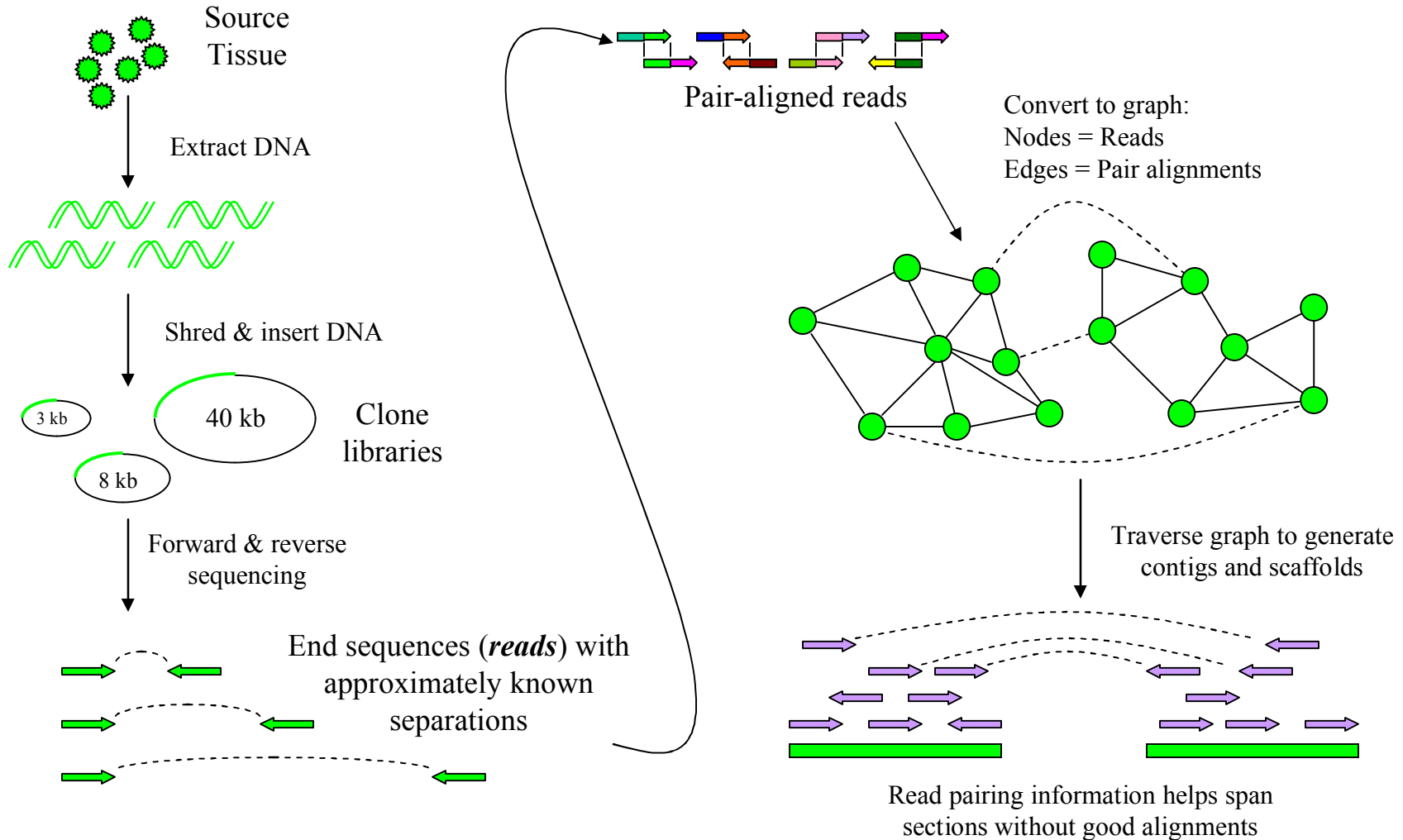
Phylogeny of eukaryotes (Baldauf, Science, 2003)



Animal (metazoan) genomes at JGI



Outline of the Assembly Process: JAZZ, the JGI In-House Assembler



Fugu Assembly Statistics

	v.2 (Aparacio et al 2002)	v.4 (Oct, 2004)
Coverage	5.4X	8.5X
Number of used reads	3,71 mln	6,92 mln
Scaffold Total	12,381	7,213
Scaffold Sequence Total	322.5 MB	393.3 MB
Scaffold N50	745 scaffolds account for 35% of sequence	125
Scaffold L50		858KB
Contig Total	45024	31213
Contig Sequence Total	332.5 MB (3% gap)	351.2 MB (10.6% gap)

Annotation Pipeline

Genome assembly

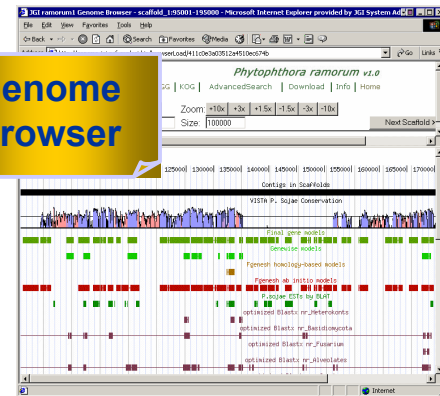
Gene prediction

Fgenesh

Genewise

Others

Genome Browser



Validation & consolidation

Homology

ESTs

Completeness

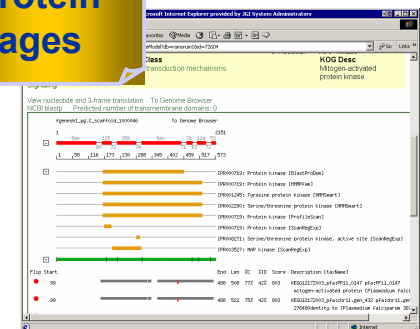
Functional Annotation

Homology

Domains

Structure

Protein Pages



Classification

GO

KEGG

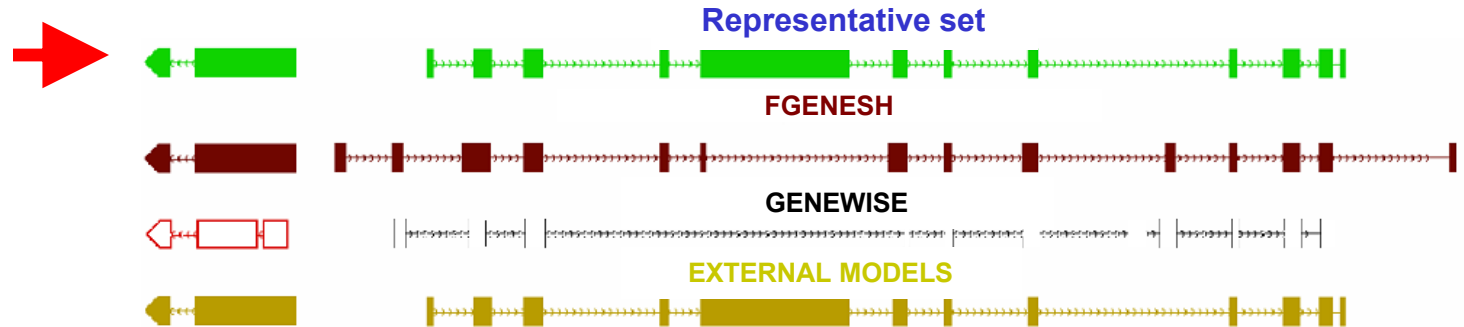
KOG

Methods used for fugu annotation at JGI

- Genewise
- Fgenesh-kg: Mapping of known genes, mRNAs, cDNAs
- Fgenesh-pm: Predictions using protein homologs
- Fgenesh-pg: Ab-initio predictions, with parameters tuned for fugu genome
- Fgenesh-est: Predictions using EST information

All predicted gene models, where possible, were EST extended into UTRs

Model selection

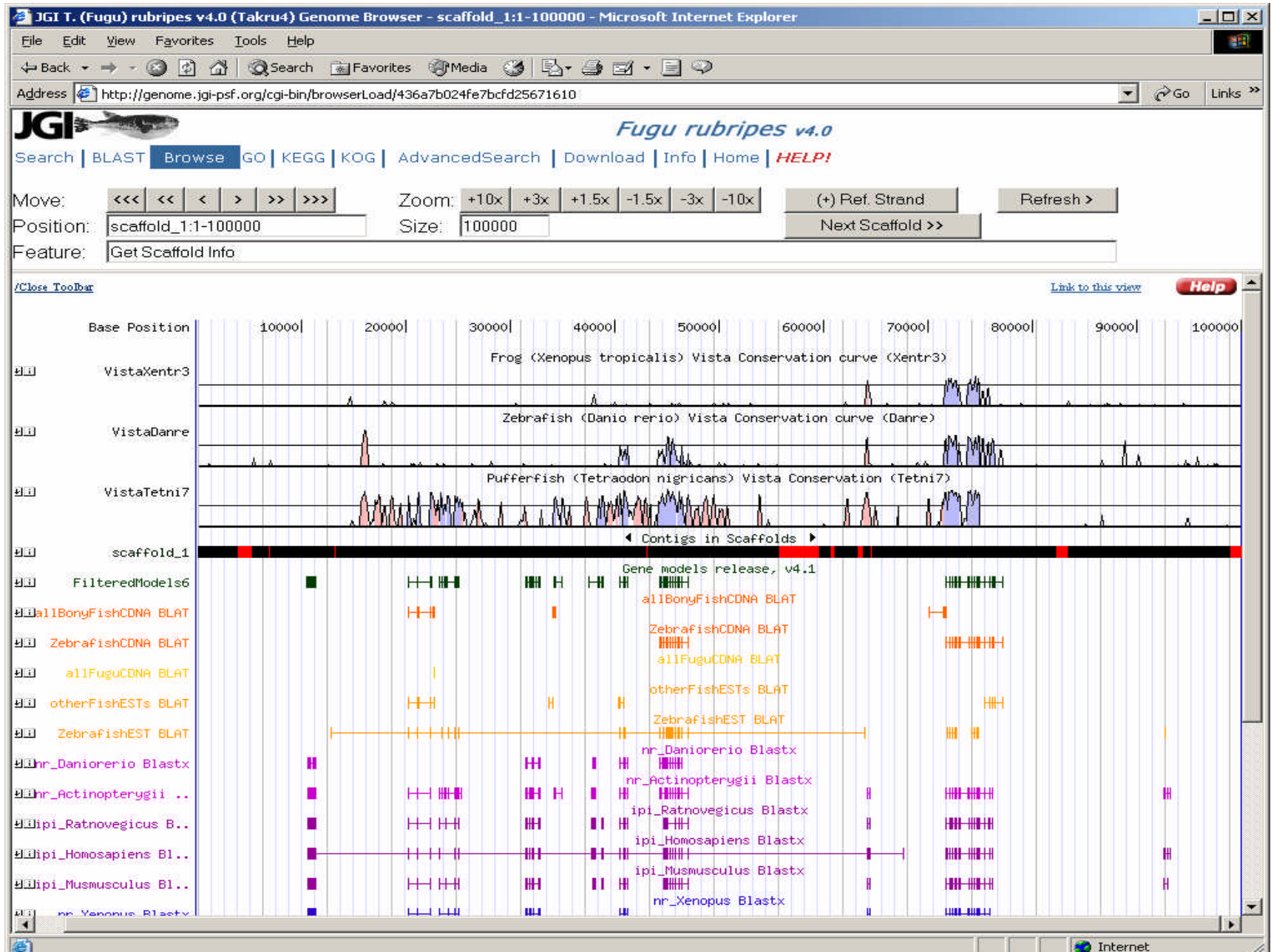


In order to produce a representative, non-redundant set of gene models, for every set of overlapping models, we select “the best” model according the following parameters:

- **Modified alignment score [$S' = S * CVR1 * CVR2$] of the best blast hit,**
- **EST coverage,**
- **Model completeness,**
- **Length of the protein/transcript**

This representative set is the first automatically generated version of a Gene Catalog, which the can be modified manually.

Genome Browser



Fraction of conserved sequence in Fugu relative to Tetraodon and Zebrafish (according to Vista)

Similarity	Tetraodon	Zebrafish
>70%	85.3 Mb (22%)	7.2 MB (1.8%)
>80%	49.1 Mb (12.5%)	2.3 Mb (0.6%)
>90%	15.6 Mb (4%)	0.2 Mb (0.06%)

Genome Browser/ProteinPages

JGI T. (Fugu) rubripes v4.0 (Takru4) Genome Browser - scaffold_1:1-100000 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail News

Address [http://genome.jgi-psf.org/](#)

JGI Search | BLAST | Browse | GO | KEGG | KOG | Adv

<<< << < > >> >>>

[/Close Toolbar](#)

scaffold_1

FilteredModels6

allBonyFishCDNA BLAT

ZebrafishCDNA BLAT

allFuguCDNA BLAT

otherFishESTs BLAT

ZebrafishEST BLAT

nr_Danioerio Blastx

nr_Actinopterygii ..

ipi_Ratnovegicus B..

ipi_Homosapiens Bl..

ipi_Musmusculus Bl..

nr_Xenopus Blastx

JGI Phytophthora ramorum v1.0 Browse - Microsoft Internet Explorer provided by JGI System Administrators

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail News

Address <http://genome.jgi-psf.org/cgi-bin/dispGeneModel?db=ramorum1&id=72604>

KOG GROUP KOG Id KOG Class KOG Desc

Cellular KOG0660 Signal transduction mechanisms Mitogen-activated protein kinase

Processes And Signaling

View nucleotide and 3-frame translation To Genome Browser

NCBI blastp Predicted number of transmembrane domains: 0

fgenesht1_pg.C_scaffold1_1000046 To Genome Browser

1 540 125 259 540 78 124 53 2151

1 58 116 173 230 288 345 402 459 517 573

IPR000719: Protein kinase [BlastProDom]

IPR000719: Protein kinase [HMMSPfam]

IPR001245: Tyrosine protein kinase [HMMSmart]

IPR002290: Serine/threonine protein kinase [HMMSmart]

IPR000719: Protein kinase [ProfileScan]

IPR000719: Protein kinase [ScanRegExp]

IPR008271: Serine/threonine protein kinase, active site [ScanRegExp]

IPR003527: MAP kinase [ScanRegExp]

Flip Start End Len %C %ID Score Description [taxName]

99 488 508 77% 42% 803 KEGG12172003_pfa:PF11_0147 pfa:PF11_0147 mitogen-activated protein [Plasmodium falci

99 488 522 75% 42% 803 KEGG12172003_pfa:chr11.gen_432 pfa:chr11.gen_276480identity to [Plasmodium falciparum 3D7

ipi_Homosapiens Blastx

ipi_Musmusculus Blastx

nr_Xenopus Blastx

Protein Pages

KOG/GO/KEGG

KOG Classification

KOG

SEARCH BY KOG ID SEARCH BY KOG KEYWORD

CELLULAR PROCESSES AND SIGNALING

<input type="checkbox"/>	M	Cell wall/membrane/envelope biogenesis	32 gene models
<input type="checkbox"/>	N	Cell motility	4 gene models
<input type="checkbox"/>	O	Posttranslational modification, protein turnover, chaperones	374 gene models
<input type="checkbox"/>	T	Signal transduction mechanisms	240 gene models
<input type="checkbox"/>	U	Intracellular trafficking, secretion, and vesicular transport	204 gene models
<input type="checkbox"/>	V	Defense mechanisms	14 gene models
<input type="checkbox"/>	W	Extracellular structures	3 gene models
<input type="checkbox"/>	Y	Nuclear structure	36 gene models
<input type="checkbox"/>	Z	Cytoskeleton	104 gene models
Total Gene Count			1011

INFORMATION STORAGE AND PROCESSING

METABOLISM

KEGG

5.4.2.2

2.7.1.2

2.7.1.63

2.7.1.2

2.7.1.63

3.1.3.11

4.1.2.13

5.3.1.1

☐ [D] GO:0003673 Gene_Ontology (Poptr1:24206)

☐ [D] [P] GO:0003674 molecular_function (22998)

☐ [D] [I] GO:0003824 catalytic activity (13605)

☐ [D] [I] GO:0016740 transferase activity (5037)

☐ [D] [I] GO:0016772 transferase activity, transferring phosphorus-containing groups (3341)

☐ [D] [I] GO:0016773 phosphotransferase activity, alcohol group as acceptor (2691)

☐ [D] [I] **GO:0004672 protein kinase activity (2461)**

☐ [D] [I] GO:0004674 protein serine/threonine kinase activity (2048)

☐ [D] [I] GO:0004712 protein threonine/tyrosine kinase activity (631)

☐ [D] [I] GO:0004673 protein-histidine kinase activity (77)

☐ [D] [I] GO:0004713 protein-tyrosine kinase activity (1955)

☐ [D] [I] GO:0019199 transmembrane receptor protein kinase activity (631)

GO

Summary of JGI annotations for Fugu v.4

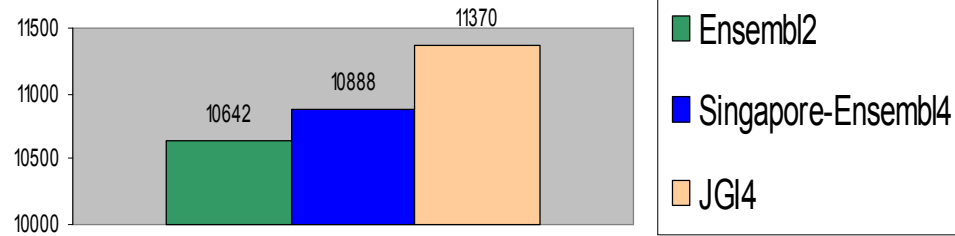
Number of gene models	26721
Number of gene models with UTR	2757 (~10%)
Average number of exons per gene	7.8
Average gene size (bp)	5690
Average CDS size(bp)	1345
Average exon size (bp)	172

JGI annotations were compared with IMCB Singapore Annotations v.4 (22,008 genes) and Ensembl annotations v.2 (20,706 genes)

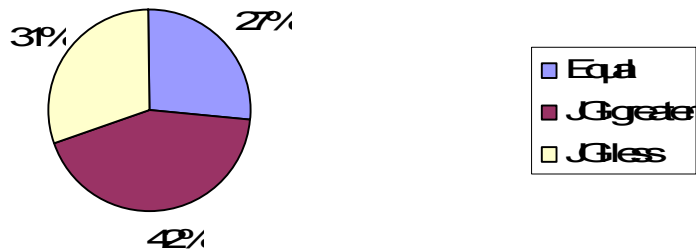
Comparison between JGI4 and Singapore4 annotations

- Total number of gene models: 26721 (JGI), 22008 (Singapore)
- ~10% (2228) of predicted genes exactly coincide
- ~26% (5633) of predicted genes are coincide without considering terminal exons
- ~70% (105,673) of internal exons coincide
- From 5143 gene models missed by Singapore, but predicted by JGI – 1361 contain some Pfam domains
- From 1107 gene models missed by JGI, but predicted by Singapore – 336 have some Pfam domains (but many of them related to repeats)

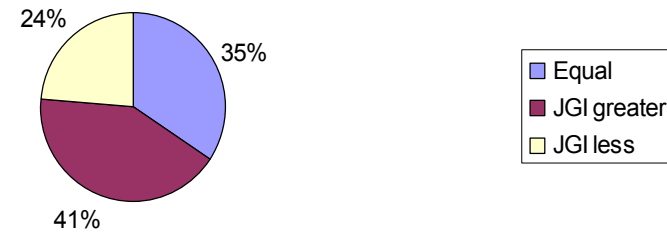
Number of putative ortologs, based on reciprocal Blastp best hits,
between Fugu and Human annotations (22,218 gene models from
Ensembl release 35)



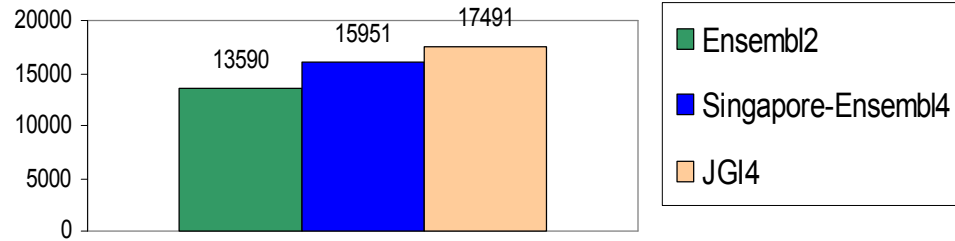
Comparison of BLASTP scores of Ensembl2 and
JGI4 on 980 common 'ortologous' pairs from Human



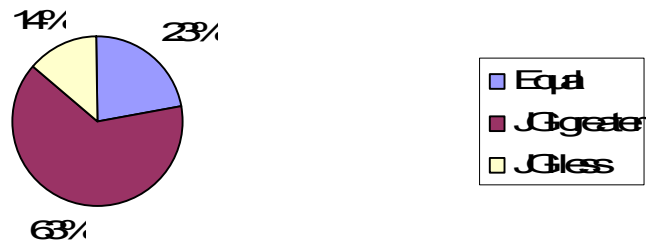
Comparison of BLASTP scores of Singapore4 and
JGI4 on 10268 common 'ortologous' pairs from
Human



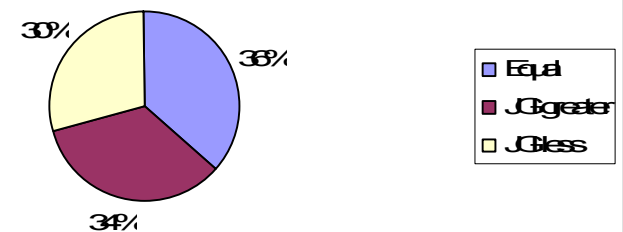
Number of putative orthologs, based on reciprocal Blastp best hits,
between Fugu and Tetraodon annotations (28,005 gene models
from Ensembl Tetraodon7)



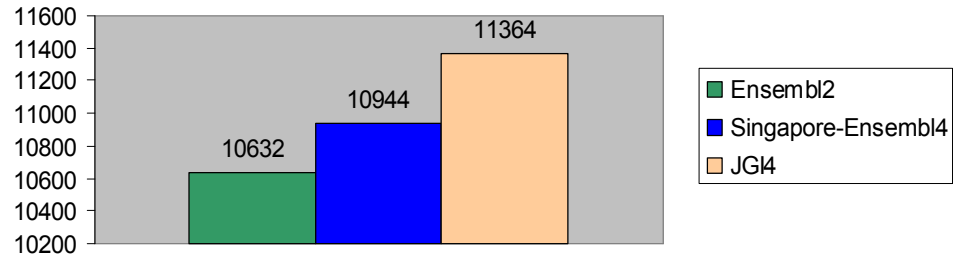
Comparison of BLASTP scores of Ensembl2 and
JGI4 on 12,831 common orthologous pairs from
Tetraodon



Comparison of BLASTP scores of Singapore4
and JGI4 on 14,776 common orthologous
pairs from Tetraodon



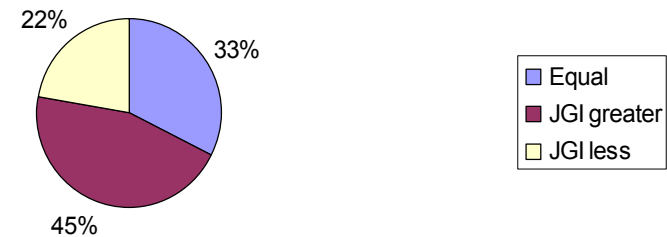
**Number of putative ortologs, based on reciprocal
Blastp best hits, between Fugu and Zebrafish
annotations (22,877 gene models from Ensembl)**



**Comparison of BLASTP scores of Ensembl2 and
JGI4 on 970 common 'ortologous' pairs from
Zebrafish**



**Comparison of BLASTP scores of Singapore4 and
JGI4 on 10178 common 'ortologous' pairs from
Zebrafish**



Pfam domain comparison between different annotations

JGI4 vs. Ensembl2

- 1546 domains occur at equal numbers in both sets
- 981 domains occur more frequently at JGI annotations (from them 145 are uniquely)
- 180 domains occur more frequently at Ensembl2 (from them 44 uniquely)

Pfam domain comparison between different annotations

JGI4 vs. Ensembl2

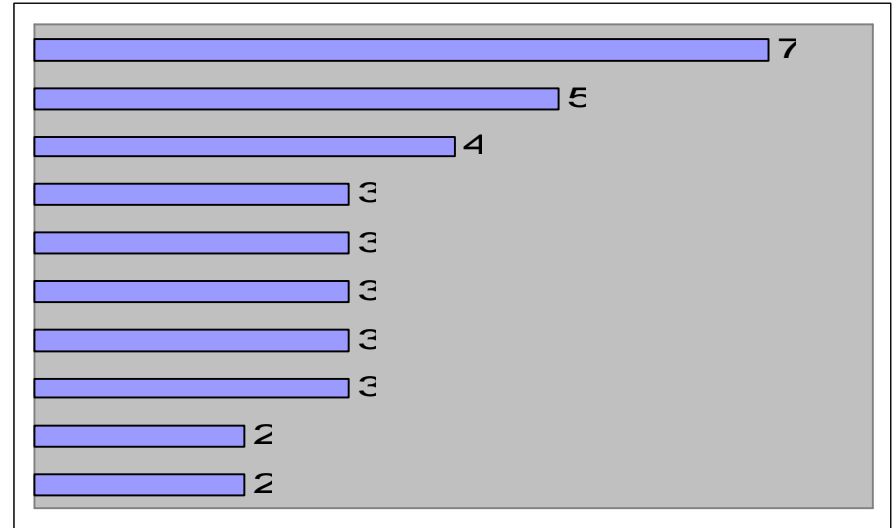
- 1546 domains occur at equal numbers in both sets
- 981 domains occur more frequently at JGI annotations (from them 145 are uniquely)
- 180 domains occur more frequently at Ensembl2 (from them 44 uniquely)

JGI4 vs. Singapore4

- 1788 domains occur at equal numbers in both sets
- 723 domains occur more frequently at JGI annotations (from them 109 uniquely)
- 195 domains occur more frequently at Singapore4 annotations (from them 43 uniquely)

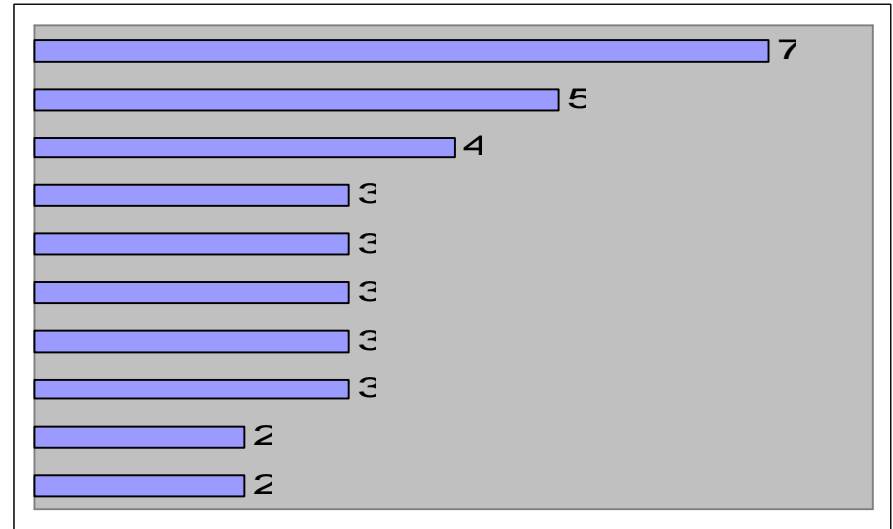
10 Most frequent PFAM domains absent in Ensembl2 annotations

uPARLy6cbnain
FAADDHMPyrindbnain
Raphilin3Aeffector chain
Endlysintypecalciumbindingrepeat (2copies)
NAD-Ubiquinoldestoquinone(complex I), various chains
Vertebrateendogenousopioidneuropeptide
EESSnolif
Poly(ADP-ribose)glycohydrolase(PARG)
NUMBphenylalaninerichregion
Autophagocytosisassociatedprotein, Cteninind chain



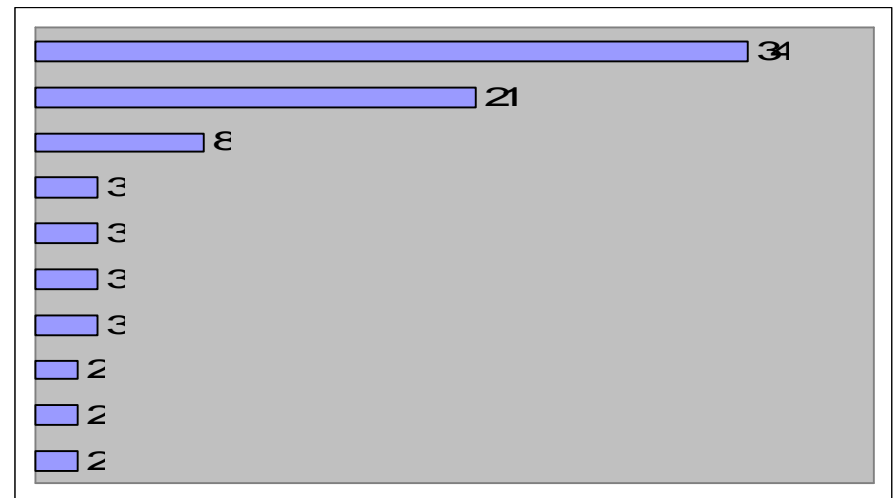
10 Most frequent PFAM domains absent in Ensembl2 annotations

uPARLy6cbnain
FAADDHMPyrindbnain
Raphilin3Aeffectorcbnain
Hebrysin type calcium binding repeat (2 copies)
NADHUbiquinone dehydrogenase (complex I), various chains
Vertebrate endogenous opioid neuropeptide
EESSnolif
Poly(ADP-ribose) glycohydrolase (PARG)
NUMBphenylalanine rich region
Autophagocytosis associated protein, Cteninid cbnain



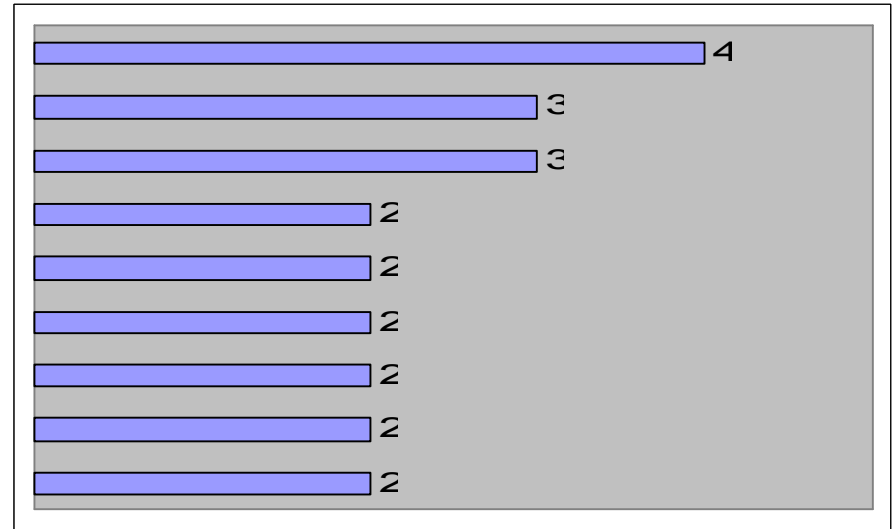
10 Most frequent PFAM domains absent in JGI annotations

Integrase corecbnain
Reverse transcriptase (RNA dependent DNA polymerase)
Retroviral capsid pcdase
CB pcdonogreNterinus, S22likecbnain
CB pcdonogreNterinid cbnain1
Transposase
CB pcdonogreNterinus, EFhandlikecbnain
Methionine
Retrotransposon gag pcdain
PcdanineP1



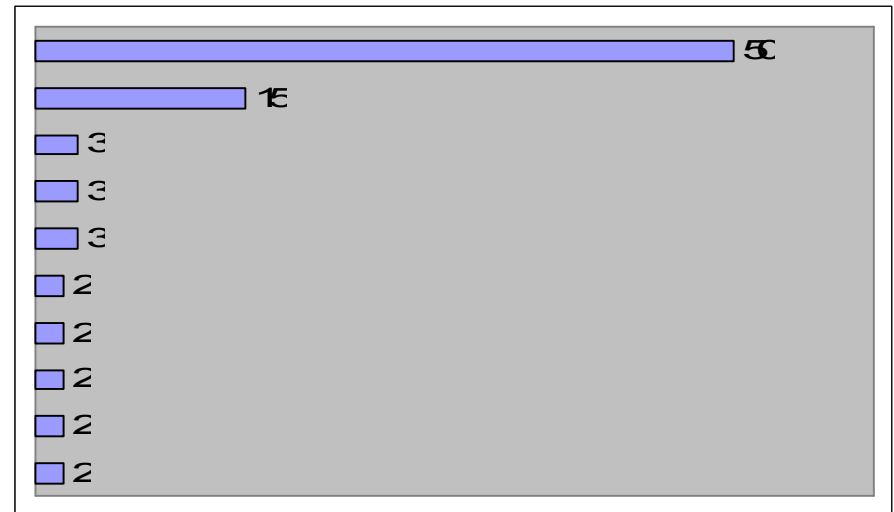
10 Most frequent PFAM domains absent in Singapore annotations

Tyrosin beta4 family
Endo-syn type calcium binding repeat (2 copies)
NADH-ubiquinol dehydrogenase (complex I), various chains
Mitochondrial-associated glycoprotein (MAGP)
Leishmanidin
CYTb domain
tRNA intron endonuclease, catalytic C-terminal domain
GLT repeat (6 copies)
Bombesin-like peptide
Transcription initiation factor II, beta subunit

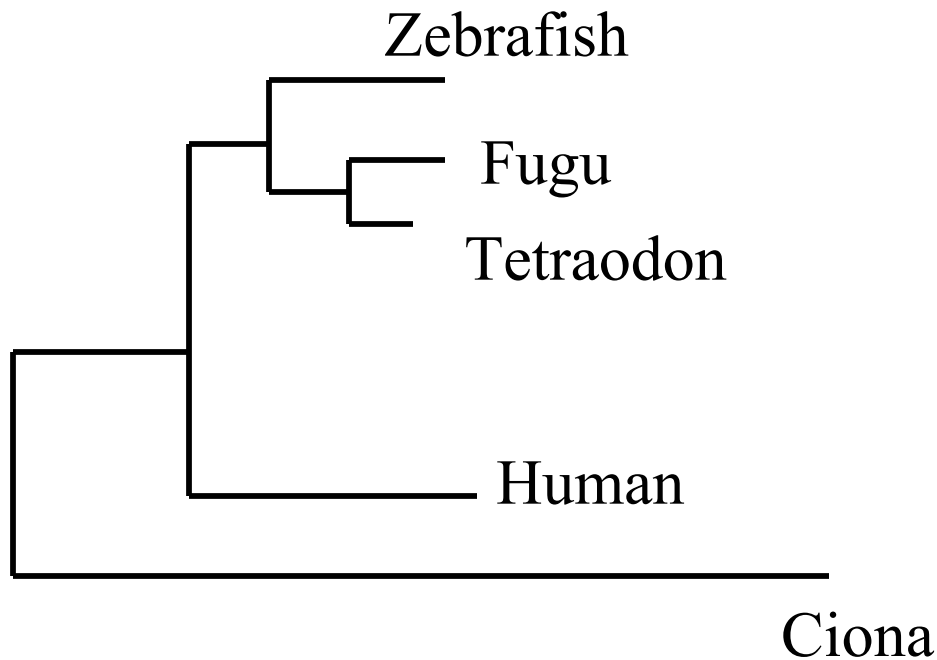


10 Most frequent PFAM domains absent in JGI annotations

Helicase core domain
Reverse transcriptase (RNA dependent DNA polymerase)
CBP domain core Neminus, S-2 like domain
CBP domain core Neminal domain 1
CBP domain core Neminus, EF hand like domain
Phosphonate kinase
Transposase
ES family
EF zinc finger
Retroviral capsid protease



NJ-tree (Phylip) based on Pfam domain content distances



FR					
TN	37				
DR	56	55			
HS	73	70	75		
CI	138	121	132	128	
	FR	TN	DR	HS	CI

Detection of possibly ‘missed’ genes based on Pfam domain profile

5 domains are in fugu (FR), but present in human(HS), Zebrafish (DR), tetraodon (TN) and ciona (CI)

HS	DR	FR	TN	CI	
1	2	0	1	1	PF07572 BCNT Bucentaur or craniofacial development
1	1	0	2	1	PF01263 Aldose_epim Aldose 1-epimerase
1	2	0	1	1	PF07065 D123 D123
1	1	0	1	1	PF02733 Dak1 Dak1 domain
1	1	0	1	1	PF01876 RNase_P_p30 RNase P subunit p30

76 domains are absent in tetraodon (TN), but present in human(HS), fugu (FR) and ciona (CI)

HS	DR	FR	TN	CI	
2	3	4	0	1	PF03332 PMM Eukaryotic phosphomannomutase
2	2	2	0	4	PF03015 Sterile Male sterility protein
2	3	3	0	1	PF05915 DUF872 Eukaryotic protein of unknown function (DUF872)
3	2	2	0	1	PF05207 zf-CSL CSL zinc finger
2	2	2	0	1	PF01127 Sdh_cyt Succinate dehydrogenase cytochrome b subunit
2	3	1	0	1	PF02936 COX4 Cytochrome c oxidase subunit IV
3	2	1	0	1	PF03943 TAP_C TAP C-terminal domain
1	1	1	0	3	PF03199 GSH_synthase Eukaryotic glutathione synthase
2	2	1	0	1	PF01765 RRF Ribosome recycling factor
1	1	1	0	3	PF05724 TPMT Thiopurine S-methyltransferase (TPMT)
2	1	1	0	1	PF01025 GrpE GrpE
1	2	1	0	1	PF05024 Gpi1 N-acetylglucosaminyl transferase component (Gpi1)
1	2	1	0	1	PF02348 CTP_transf_3 Cytidylyltransferase
1	1	2	0	1	PF05060 MGAT2 N-acetylglucosaminyltransferase II (MGAT2)
1	2	1	0	1	PF02441 Flavoprotein Flavoprotein

156 domains are absent in zebrafish (DR), but present in human(HS), fugu (FR) and ciona (CI)

HS	DR	FR	TN	CI	
11	0	2	2	2	PF03184 DDE DDE superfamily endonuclease
4	0	5	4	2	PF06814 Lung_7-TM_R Lung seven transmembrane receptor
3	0	4	4	3	PF04675 DNA_ligase_A_N DNA ligase
3	0	3	3	3	PF04567 RNA_pol_Rpb2_5 RNA polymerase Rpb2
2	0	4	3	3	PF00464 SHMT Serine hydroxymethyltransferase
2	0	3	3	2	PF02516 STT3 Oligosaccharyl transferase STT3 subunit
3	0	3	3	1	PF03836 RasGAP_C RasGAP C-terminus
2	0	2	2	3	PF02010 REJ REJ domain
2	0	2	1	3	PF03942 DTW DTW domain
2	0	2	2	2	PF02515 CoA_transf_3 CoA-transferase family III
2	0	2	2	2	PF01142 TruD tRNA pseudouridine synthase D (TruD)
2	0	3	2	1	PF05822 UMPH-1 Pyrimidine 5'-nucleotidase (UMPH-1)
2	0	2	2	2	PF02969 TAF TATA box binding protein associated factor (TAF)
1	0	2	1	2	PF01464 SLT Transglycosylase SLT domain
1	0	1	1	2	PF05090 VKG_Carbox Vitamin K-dependent gamma-carboxylase

Acknowledgements

JGI:

Nik Putnam

Astrid Terry

Igor Grigoriev

Dan Rokhsar

IMCB (Singapore)

Eddie Loh

Burappa Venkatesh

Ensemble team: (for providing data on different annotations)